APPENDIX D
MAP

SCORE USE, MEANINGFULNESS, AND DEPENDABILITY

The Missouri Assessment Program: Score Use, Meaningfulness, and Dependability

The Missouri Assessment Program (MAP) is one of several educational reforms mandated by the Outstanding Schools Act of 1993. As a result of this legislation, the State Board of Education directed the Missouri Department of Elementary and Secondary Education (DESE) to identify the knowledge, skills, and competencies that Missouri students should acquire by the time they complete high school and to assess student progress toward these academic standards. DESE staff worked with educators, parents, and business professionals from throughout the state to develop the Show Me Standards and to create the MAP as a tool for evaluating the proficiencies represented by the Standards.

The MAP currently includes mathematics assessments for grades 4, 8, and 10; communication arts assessments for grades 3, 7, and 11; science assessments for grades 3, 7, and 10; social studies assessments for grades 4, 8, and 11.

Each MAP assessment requires about three hours of testing time, and most assessments include three types of test items: multiple choice, constructed response, and performance events. For most assessments, the multiple-choice component is the survey portion of the *Terra Nova*, a nationally normed achievement test published by DESE's MAP contractor, CTB McGraw-Hill. (The social studies assessments include multiple-choice items that are not from the *Terra Nova* and the health/physical education and fine arts assessments do not utilize any *Terra Nova* items, although they do contain multiple-choice items.) Constructed-response items require students to supply an appropriate answer and, in some instances, to show their work. Performance events call for students to work through more complicated problems and may allow for more than one approach to arrive at a correct answer. All three of these item formats, but especially the latter two, require students to apply what they have learned to complex, real-life situations.

Appropriate uses of MAP scores

MAP scores provide information about what individual students know and can do relative to the Show-Me Standards. For individual students, DESE and CTB report a MAP scale score, a MAP achievement level, and a *Terra Nova* national percentile. Educators may use these quantitative and qualitative results to make inferences about student's proficiency relative to the content and process Standards assessed at that grade and subject.

Educators and policy makers may appropriately use MAP results for groups of students to judge the effectiveness of educational programs and services offered at the local level. DESE uses group-level data from the MAP in the Missouri School Improvement Program review process, and DESE encourages district personnel to use these scores to conduct their own internal evaluations, to monitor progress over time, and to inform planning for the future. DESE also

uses data from MAP administrations to report to the public about the quality of education in the state.

Judging the quality of assessment results

When we judge assessment results, we must consider two important qualities—how meaningful or "valid" the results are for their intended purpose(s) and how dependable or "reliable" the results are. These two characteristics are closely connected; in fact, score dependability limits score meaningfulness. We can evaluate assessment data by examining score dependability, but we must also consider score meaningfulness if we want to arrive at sound judgments, about the worth of results.

Meaningfulness or "validity" of MAP scores

First and foremost, we ensure the meaningfulness or validity of MAP scores as indices of proficiency relative to the Show-Me Standards by using methodical and rigorous test-development procedures. CTB and DESE have developed MAP assessments in accordance with accepted procedures and criteria (as articulated, for example, in Standards for Educational and Psychological Testing, AERA, APA, NCME, 1985), intentionally aligning MAP assessments to the specific Show-Me Standards being measured at that grade and subject area. For each assessment, content experts determined that the *Terra Nova* items for that grade and subject measure the Standards, and Missouri educators wrote constructed-response items and performance events that match the designated Standards. Then, groups of Missouri educators reviewed each item to insure that it did indeed measure the content or process called for in the Standard. The "item-to-Standard" congruence ratings that these reviewers produced provide evidence for the meaningfulness of MAP scores.

Another way to verify the meaningfulness of MAP scores is to investigate the underlying psychological traits or "constructs" that a given assessment measures. CTB and DESE routinely examine how performance on individual items related to performance on other items and how performance on an individual item relates to performance on the entire assessment. The various item- and score-pattern analyses conducted on MAP results show that each assessment is measuring the traits it is intended to measure (e.g., communication arts assessments measure reading and writing skills) and does not measure unrelated constructs.

A third type of evidence supporting the meaningfulness of MAP results comes from a recent study of the "consequential validity" of the MAP. This research, conducted in 1999 by the Center for Learning, Evaluation, and Assessment Research at the University of Missouri-Columbia, investigated the consequences resulting from the implementation of the MAP, focusing specifically on changes in instructional practices in mathematics. Researchers concluded that changes are occurring, primarily in the area of teacher beliefs and perceptions. Study results indicated that teachers are becoming more convinced of the work of authentic learning activities and assessment methods. In addition, researchers found that

teachers are revising their grading practices as a result of the MAP, using more performance-based methods to determine grades than in the past.

The process of collecting evidence for the meaningfulness of assessment data is ongoing, as is the process of ensuring meaningfulness through sound test-development procedures. CTB and DESE will continue to conduct validity studies on future editions of the MAP and to build meaningfulness into results by adhering to industry standards during test-development stages. However, we have very firm evidence that the MAP assessments yield scores that are valid, given the stated purposes of the program. Scores provide information about students' attainment of the Show-Me Standards and can be appropriately used to fulfill the charges stipulated in the Outstanding School Act.

Dependability or "reliability" of MAP scores

We build score dependability or reliability into the test-construction process, just as we do score meaningfulness. We know that all educational test scores reflect some degree of error; no mental measurement is perfect. We also know that error can come from a variety of sources: the instrument itself, the examiner, the assessment environment the scoring process, and, in the case of assessments like the MAP, in the process of establishing cut-point scores for the various achievement levels. How much error are we willing to tolerate? The answer to this question varies depending on the purpose of assessment. Scores that are used to make high-stakes decagons for individuals must be more dependable than scores that are used to make decisions of lesser import or judgments that pertain to groups of students.

Developers of educational assessments make every effort to create high-quality instruments that will yield dependable (and, of course, meaningful) scores. In an assessment program like the MAP, which includes constructed-response items and performance events that must be scored by knowledgeable scorers (as contrasted to selected-response items that can be scored by a machine using a key), developers also go to great lengths to ensure that the scoring process yields consistent information. CTB and DESE have put stringent procedures in place to ensue reliable scoring of MAP items.

**Dependability of scale scores*

Score dependability or reliability can be quantified and reported as a number ranging from 0 to 1; the higher the coefficient, the more dependable the score. Table 1 presents reliability coefficients for MAP assessment scale scores for every operational year. All coefficients are high and indicate that we can have confidence in MAP scale scores. (It is important to keep in mind that it is these overall l scale scores for each assessment that are used for decision-making purposes.)

**Dependability of scores from open-ended items*

While we appropriately place primary emphasis on the overall reliability of a given MAP assessment score, we also have to consider the dependability of the scores derived from the subset of items that are judged by human readers—constructed-response questions and performance events. We know that we lose a bit of reliability when we use open-ended items that cannot be scored by a machine. However, what we lose in reliability, we gain in meaningfulness or validity –these types of items are much more representative of "real life" than multiple-choice items. (And, given the reliability coefficients for MAP scores, it is clear that we are not losing much in the way of dependability.)

One way to think about the dependability of open-ended item scores is to consider the percent of perfect agreement—the percent of cases for which two readers assign the same response the same score. Table 2 shows the median percent of perfect agreement for the 1999 and 2000 MAP assessments. These indices range from 75% to 96% and suggest that scorers are reaching perfect agreement much of the time.

Yet another way to think about the dependability of open-ended item scores is to consider the percent of adjacent agreement—the percent of cases for which two readers assign scores that are adjacent to (within one point of) one another. When adjacent agreement is used as the basis for defining reliability, percents of agreement are much higher; in fact, most of these indices are well above 95%. For example, on the 1999 communication arts grade-7 assessment, the percent of adjacent agreement ranged from 95% to 100%, with the median percent equal to 99%. On the 1999 mathematics grade-10 assessment, the percent of adjacent agreement ranged from 92% to 100%, with the median percent equal to 98%.

**Dependability of achievement-level classifications*

CTB and DESE use the "bookmark procedure" to set achievement levels (step 1, progressing, nearing proficiency, proficient, advanced) for the MAP assessments. This approach, which is described below, has been successfully applied to a number of state assessment program achievement-level settings.

In this procedure, standard-setting panels are given booklets with items ordered according to level of difficulty. Panelists study the ordered item booklets to observe the increase in the knowledge, skills, and abilities required of students as the items increase in difficulty. They examine each item, discussing what the item measures and why it is more difficult than preceding items in the booklet. Each panelist moves through the ordered item booklet until the level of item difficulty surpasses that which the given achievement-level students (e.g. proficient) should be expected to know and be able to do (based on the panelists' expectations and the Show-Me Standards). Each panelist places a bookmark at this position in the ordered item booklet. One bookmark, is placed for each of the required cut points. Items preceding participants' bookmarks reflect content that all students at the given achievement level are expected to know and be able to

do (according to panelists' expectations). Several rounds of judgments occur, and cut points are ultimately determined that translate the panelists' expectations into appropriate achievement levels.

The bookmark procedure incorporates expert judgments as well as empirical data into the achievement-level setting process, so it builds a great deal of information into the cut-point scores. However, like examinees' scores, cut-point scores reflect some degree of error. Nevertheless, a careful inspection of reliability data for the achievement-level cut points indicates that the achievement-level classifications are highly reliable, especially given that panelists are creating five categories of performance. The degree of error associated with the cut-point scores is very low, ranging from 1 to 6 scale score points across the mathematics, communication arts, and science assessments.

**Comparing MAP reliability data to data from other tests*
It is worthwhile to compare these MAP coefficients with reliability data for other tests of a similar nature; Table 3 presents a sampling of this sort of data. Such a comparison shows that the dependability of MAP scores compares quite favorably with the reliability of scores from other well respected instruments that incorporate open-ended as well as selected-response items.

Conclusion

There is ample technical evidence to support the claim that MAP scores are reliable and valid measures of achievement relative to the Show-Me Standards. They are, in fact, more reliable than results from several other tests that are used of similar purposes. CTB, DESE, and Missouri educators can and should have confidence in MAP results.

References

AERA, APA, & NCME. (1985). *Standards for educational and psychological testing.* Washington, DC: Author.

Allen, M.J., & Yen, W.M. (1979). *Introduction to measurement theory*. Belmont, CA: Wadsworth, Inc.

Brown, W. (2000). *MAP achievement-level setting summary*. Unpublished manuscript, Missouri Department of Elementary and Secondary Education.

CTB McGraw-Hill. (1999). *Missouri Assessment Program: Guide to test interpretation*. Monterey, CA: Author

Table 1

MAP Scale Score Reliability Coefficients

	1997	1998	1999	2000
Mathematics				
Grade 4	.919	.921	.915	.913
Grade 8	.931	.927	.927	.929
Grade 10	.936	.940	.929	.940
Communication Arts				
Grade 3		.920	.915	.913
Grade 7		.932	.905	.907
Grade 11		.939	.919	.917
Science				
Grade 3		.907	.903	.903
Grade 7		.915	.875	.918
Grade 10		.916	.908	.882
Social Studies				
Grade 4			.918	.923
Grade 8			.906	.921
Grade 11			.925	.885

Table 2

Median Percent of Perfect Agreement

	1999	2000
Math		
Grade 4	95.64	96.04
Grade 8	89.04	92.33
Grade 10	84.39	89.39
Communication Arts		
Grade 3	82.94	84.12
Grade 7	71.40	88.20
Grade 11	75.00	77.12
Science		
Grade 3	88.81	92.64
Grade 7	82.03	86.56
Grade 10	81.85	87.13
Social Studies		
Grade 4	78.78	70.93
Grade 8	78.04	75.00
Grade 11	75.16	78.79

Table 3

Reliability Information for Educational Assessments Similar to MAP

Stanford Achievement Test, 9th Edition

Mid .80's to .90's for entire test
.60's to low .80's for open-ended assessments

Advanced Placement Examinations

	Composite Score	Open-Ended Items
U.S. History	.88 to .92	.62 to .78
Biology	.93 to .96	.72 to .89
Chemistry	.94 to .98	.86 to .98
English/Language/Composition	.85 to .88	.67 to .76
U.S. Government	.87 to .93	.67 to .87

SAT I

Verbal	.91 to .93
Math	.92 to .93

SAT II

	Composite	Essay
Writing	.86 to .91	.58

ACT Assessment

English	.90 to .91
Mathematics	.89 to .91
Reading	.86 to .87
Science Reasoning	.82 to .86

MMAT

	Reading	Math	Science	Social Studies
Grade 2	.95	.85		
Grade 3	.94	.92	.90	.92
Grade 4	.95	.91	.89	.93
Grade 5	.94	.93	.90	.94
Grade 6	.93	.96	.93	.95
Grade 7	.95	.93	.89	.95
Grade 8	.95	.94	.88	.94
Grade 9	.95	.94	.88	.94
Grade 10	.95	.93	.91	.95